*Proceedings of the IEOM International Conference on Smart Mobility and Vehicle Electrification*
*Detroit, Michigan, USA, October 10-12, 2023*

# A Preliminary Study on Human Trust in Pseudo-Real-Time Scenario through Electroencephalography and Machine Learning based Data Classification

**Kazi Farzana Firoz, Younho Seong, Sun Yi, Yoo-Sang Chang**
North Carolina Agricultural and Technical State University
Greensboro, NC 27411, USA
kfiroz@aggies.ncat.edu, yseong@ncat.edu, syi@ncat.edu, ychang@aggies.ncat.edu

## Abstract

This study aims to sense trust and distrust in a real-time inspired scenario through the classification of brain signals. Here, a word elicitation study is used to invoke the mental state associated with trust and distrust associated with the machine. Participants think of any event or experience that comes into their mind when they observe the word. They think or recall that event/experience without deliberately filtering out any kind of cognitive or affective mental state, which we consider as a replica of a real-life scenario where all kinds of mental states or emotions possibly co-exist along with trust or distrust. While thinking or recalling such events, Electroencephalography data is recorded from the participants' cortex and analyzed through Machine Learning approaches with several classification algorithms. The study developed an approach to sense whether the human is going through trust or distrust and compared different methods to discuss their efficacies in different scenarios. Here, the individualistic and generalistic approach is delved into, and it found that individualistic approaches provide better accuracy in sensing trust or distrust state of the human brain. Also, this study explored ways to increase the efficiency of the method by reducing the number of channels and compared the performance of the models by observing the loss of accuracy caused by the reduced number of channels. This study found that the K-Nearest Neighbor and/or Random Forest classifier algorithm provides the best result using raw data with the individualistic approach in most scenarios, achieving up to 100% average accuracy.

## Keywords
Trust, Human-Machine systems, Electroencephalography (EEG), Neural Signal and Machine Learning

## 1. Introduction
Trust is an important factor in everyday life whenever we perform any tasks which might include some level of uncertainty and risk. It is a crucial factor in any teamwork, whether it be human-human teaming, human-machine teaming, or even human-animal teaming. A horse will not cooperate with its human rider if it does not trust him. Similarly, a human would not use a machine if they do not have proper confidence in the machine. And here comes the issue of mistrust, under-trust, and over-trust (Parasuraman and Riley 1997). Especially with the development of technology and intelligent systems, it has become more crucial as intricacy and high level of technology make it unfamiliar to the user. Have you ever seen an elderly person having a smartphone but not using all those smart features just because he does not have confidence in them? This is an example of undertrust. Also, sometimes, we install different apps or go to different websites and click the button "I agree" without reading the terms and conditions properly; this is an example of overtrusting the app or website. There are also more advanced technological scenarios where human-machine team bonding needs to be fine-tuned, such as operation theatre and warzone; this tuning can be done either by providing proper training to the human or by adjusting the machine according to the human's convenience. Human-machine trust can be described in a more technological way by using Sheridan's seven characteristics of trust: reliability, robustness, familiarity, understandability, explication of intention, usefulness, and dependence (Sheridan 1989). With the advancement of technology, as the dependency of humans on machines is increasing, the machines, and technologies are becoming more and more complex; it is becoming more difficult to ensure all the characteristics, such as familiarity, understandability, and so on.

When any level of automation and intelligent system is involved, and human has to team play with that machine, the trust needs to be precisely established. As an example, we can consider a level 4 autonomous car, where the human

must know when to intervene and when to rely on the system. It is also critical to calibrate whether the human does or does not trust the system. A person can be overenthusiastic and can be careless in required situations; in the opposite way, a human can be just less confidence due to lack of familiarity, understandability and if there is no clarity of machine's intention; and can just intervene in the autonomy action where s/he does not need and thus can lead to less performance of the car. Sometimes, human users cannot be clear whether they are trusting or not the advanced machine. Lots of time, they just feel "somewhat" trust or distrust. But sensitive situations like warzone and operation theatre where lives are at stake do not offer places for feeling "somewhat." Also, trust is a dynamic component in the context of human-machine teamwork, which can vary from time to time and thus can alter the performance of the human-machine team.That is why researchers are studying trust from different perspectives and for different situations to find a measure to calibrate human trust efficiently. In the past, trust was a topic of interest confined to the psychology domain. But nowadays, with the increased importance of establishing human-machine trust, it has become the focal point in other domains, too, such as rehabilitation, Engineering, Neurology, Ergonomics, and so on. Usually, cognitive or affective aspects of the socio-behavioral construct, like trust, is investigated through qualitative study, such as survey or questionnaire. In this demanding era of human-machine trust, only qualitative study might not be enough. Also, when machines are involved, quantitative studies may have additional benefits.

Among different quantitative methods of understanding and studying the human mind, one promising method is neural signal investigation. Until a few years back, investigating neural signals would be considered as a novel method. But as technology develops, the neural signal is being studied to understand a number of human brain functions, even to study novel topics like intuition (Firoz and Seong 2023) and decision making (Firoz and Seong 2023). Among different methods of human brain imaging and neural signal observation, Electroencephalography (EEG) has become a popular method to study the human brain due to its convenient characteristics such as comparatively low cost, easily handleable, the requirement of minimum training, robustness, and satisfactory resolution. In this study, we considered a real-life-like scenario to investigate trust or distrust. This means we attempted to simulate a complex state of the human brain where trust (or distrust) will co-exist with possible other cognitive and affective functions and states such as emotion, memory, etc. We record and analyze the corresponding EEG data to identify trust or distrust state by machine-learning technique. In the following sections, we will discuss in detail the background, methodology, and results.

### 1.1 Objectives

The objective of the study is to identify trust and distrust moments in a practical situation in a moment-by-moment scenario with the most efficiency. With that goal, we developed a model with a word-elicitation study and explored different algorithms, parameters, and factors which can improve (or reduce) the efficacy of the model. We envision a human-machine team structure where the machine will read human-brain signals by receiving live data streaming, which probably will also take command from the human brain signal (like BCI wheelchair or moving a nano-bot), identify the trust/distrust state in a moment-by-moment manner (which means can sense a single changed moment of distrust/trust, and adjust its function accordingly (e.g., increase/decrease information in display, adjust alert, etc.) as well as continuously re-train the trust-sensing model to improve accuracy. The envisioned model is shown in Figure 1. In that kind of live scenario where trust/distrust can change within a single moment, we would like to have a trust/distrust identification that can perform fast, will cause minimum discomfort, and can achieve satisfactory (depending on the situation) accuracy. With this aim, we design a method to identify trust/distrust with a word elicitation study to simulate a brain situation near a real scenario of trust/distrust, where other mental states are also present.

In this preliminary study, we performed EEG data of neural signal classification of this pseudo-real trust/distrust states by using the Machine Learning technique to identify the most suitable algorithm for the model in different scenarios. We also considered the challenge that some Machine Learning algorithms may take longer time than that would be suitable for a real-live function performance. We compared several algorithms and data handling measures, identified a lower number of channels, and observed the reduced feature performance with the aim of increasing efficiency.

## 2. Literature Review

From the perspective of human-machine teaming, trust can be described scientifically with three characteristics: predictability, dependability, and faith (Barber 1983). These characteristics are quite aligned with the traditional

definition of trust that can be found in the dictionary. A more technological interpretation of human-machine trust can be considered as a function of three attributes: persistence, technical competence, and responsibility (Muir 1987). Human-machine trust can be layered into three components: dispositional, situational, and learned (Hoff and Bashir 2015). Dispositional trust is the component that comes from an individual's characteristics, such as culture, age, gender, personality, and situational, and learned trust can be understood as their name indicates. All of this interpretation of trust resonates with the seven characteristics of human-machine trust described by Sheridan (Sheridan 1989).

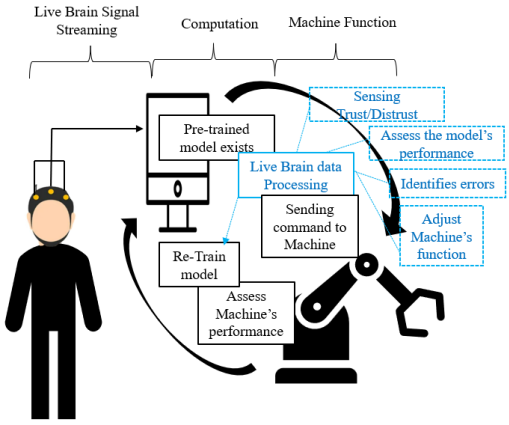



Figure 1. Human in-loop with machine with the live-sensing neural signal; the blue part emphasizes the process involved around live data

From different interpretations of trust in the context of human-machine teaming, it can be understood that a large component of human trust comes from an individual's experience. If humans experience poor performance from the machine, they will not trust the machine anymore (Oh et al. 2020). Even trust can be changed in one single interaction (Hoff and Bashir 2015). That is why trust in a human-machine context is considered a dynamic component that can be altered within a moment during a live performance in a practical situation. A study investigated this dynamic nature of trust in a live-sensing method, using EEG and GSR (Galvanic Skin Resistance), and found that customized (individual-based) features provide more accuracy than general features (Akash et al. 2018), which supports the statement that large variance of trust arises from human's individual characteristics, which not only reflects in neural signal but also in physiological (in this case Skin Resistance) features. Another study found that, even though the machine characteristics remain constant, human's individual perception can differ for the characteristics (Merritt and Ilgen 2008); they found that 52% of trust variances are influenced by an individual's perception, and trust can vary over time.

Trust can also be different in terms of award, penalty, and interest (Boudreau et al. 2009), which means trustors will always consider the value of the stake associated with their trust, and depending on that, their behavioral outcome may change. To understand trust in human-machine context, a preliminary study observed that the presence of a human-like cue of a machine does not affect the trust behavior of the human (Dong et al. 2015), which means whether it is a humanoid robot or an AI in a computer does not change trust behavior of human.

A neural investigation by EEG of trust and mistrust situation found an association between effective performance and concentration of trust and an association of stress and anxiety with mistrust situation (Boudreau et al. 2009), which means establishing proper trust will improve the performance of the human-machine team. To establish trust, we can consider the seven characteristics of Sheridan (1989): We may enhance reliability, usefulness, and robustness by improving the machine design, we can enhance familiarity through human user's training, we can arrange to present detailed information to increase understandability and explication of intention; the operator will deliberately depend more on the system if s/he can trust the system enough. The reward of all the enhancement comes with one challenge, which is efficiency; overly presentation of information can cause a higher workload. Even one study has found that higher transparency increase trust in human when the trust is low but decreases the trust when the trust of the human operator is already high in the machine (Akash et al. 2018).

Therefore, calibrating trust in human-machine is not a straightforward topic. Situational and individual components are highly correlated with trust. One study found that human-human trust has a significant difference from the general trust or human-machine trust; whether human-machine trust is quite similar to general trust (Jian et al. 2000). This study also developed a scale to measure trust in human-machine context through "word study" or "word elicitation study." This word elicitation study is helpful in the neural signal study of trust because by using this word trust or distrust state can be 'stimulated' in the human brain. Previous studies have also used this method to investigate neural signals of trust to find brain regions and to develop strategies to identify trust/distrust situations (Firoz et al. 2022). In light of the above literature, we are developing an approach by which we can identify trust or distrust scenarios given the dynamic characteristics of trust and the high individual effect on human-machine trust.

## 3. Methods

### 3.1 Electroencephalography (EEG)

We used the EEG (Electroencephalography) technique to collect and record the participant's brain signal. One of the challenges of using the EEG method is it can cause discomfort as sometimes headsets with a number of electrodes can be heavy. Keeping this in mind, here we used 30 electrodes covering all the cortex and aimed to observe the accuracy with the lower number of electrodes. In this study, an Emotiv EpocFlex device is used to collect and record neural signals. A head cap is worn by the participant, which consists of electrodes following the 10-20 system of electrode management. In this study, 30 channels have been used to collect neural signals with two more electrodes as references, as shown in Figure 3(a). The data was collected at a 128Hz sampling rate. EmotivPro software has been used to extract the brain signal data as a digital form to feed directly for classification. For data pre-processing and cleaning, we used the EEGLAB extension of MATLAB software.

### 3.2 Word Elicitation Method

The word elicitation method has been used here to invoke trust and distrust states in participants' minds. A previous study (Jian et al. 2000)has developed a scale to measure human-machine trust using words, and they identified 15 top words (among 112) having the most association with Trust and the top 15 words associated with distrust, as shown in Table 1 and Table 2. They also described the cue lines to describe the relationship between the words and machines. We have used these 30 words as presented on a computer screen in front of the participant. The words were presented accompanied by the cue lines, which would help the participants to think about the machine or system.

### 3.3 Machine Learning Technique

In this study, we used machine learning techniques and six of the classification algorithms.

**Logistics Regression (LR)** is a widely used tool for binary classification scenarios. It predicts the dependent variable outcome for given independent variable values based on the learning from the training dataset. It is suitable for binary problems, as is ours.
**Linear Discriminant Analysis (LDA)** uses the technique of feature selection and reduction before classification analysis and has the limitation of being better suitable for linearly solvable problems and normally distributed datasets. However, we still observed its performance for our datasets.

**K-Nearest Neighbor(KNN)** is a non-parametric approach that works well with binary as well as multi-class datasets and works well with noisy data. That is why we consider this algorithm highly suitable for our model, as we are considering using an un-cleaned (raw) dataset; it certainly consists of some noise.

**The Decision Tree (DT)** classifier can use different feature subsets as well as decision rules for classification. It also works well with non-linear problems, which makes it another potential good algorithm for our datasets.
**Gaussian Naïve Bayes (NB)** algorithm assumes normal or Gaussian distribution for the dataset. Still, the non-expensiveness nature encouraged us to explore this algorithm's performance.
**Support Vector Machine (SVM)** is another widely used algorithm for supervised Machine Learning problems. Problem. However, we had to carefully select the parameters to get better performance. It has a disadvantage in that it can take a longer time to train, which may make it not so suitable for a fast-need situation.
**Random Forest (RF)** is another algorithm suitable for large datasets with noise, but it takes a longer time to train; therefore, it needs to assess the context of interest and whether it would fit for the model.

### 3.4 Data Analysis

After collecting the recording, we marked the data for trust and distrust words in the associated time domain. Then, we used the data directly for classification analysis without any cleaning or artifact removal. We considered avoiding cleaning for two purposes: One, in the process of cleaning, we may lose information that may be important for achieving higher accuracy; in this study, our focus is to achieve better classification accuracy in an efficient way; thus, if some 'corrupted' component of data helps us to increase accuracy, we prefer that than 'un-corrupted' brain data which will give us not-so-high accuracy; Two, cleaning of EEG data of neural signal means removing artifacts such as eye components, line noise, heart or muscle components, etc., which requires extensive and lengthy pre-processing which contradicts our aim is to achieve accuracy in an efficient way. Still, we performed a cleaning of one participant's dataset to observe the effect. We considered trust and distrust as two classes for classification analysis and all 30 channels as individual features. In this discussion, features and channels will have the same meaning. All the sampled data in the corresponding timeframe is considered as data points. With 15 words for 30 seconds each and a 128 sampling rate, ideally, we have 57,600 data points for trust (and 57,600 data points for distrust, a total of 115,200 data points) for each participant, in some cases, due to minor connection disruption, we may have a lower number of data points, but the amount of this kind of lost data point is very small to affect the analysis.

Table 1. Trust Words (Jian et al., 2000)

| Trust words and Cues | |
|---|---|
| Assurance | I am confident in the system |
| Confidence | I am confident in the system |
| Entrust | I can trust the system |
| Familiarity | I am familiar with the system |
| Fidelity | The system is dependable |
| Friendship | The system is reliable |
| Honesty | The system is reliable |
| Honor | The system has integrity |
| Integrity | The system has integrity |
| Love | The system is reliable |
| Loyalty | The system is dependable |
| Promise | The system is reliable |
| Reliability | The system is reliable |
| Security | The system provides security |
| Trustworthy | The system is reliable |

Table 2. Distrust Words

| Distrust words and Cue | |
|---|---|
| Betray | The system is deceptive |
| Beware | I am wary of the system |
| Cheat | The system is deceptive |
| Cruel | The system's outcome will have a harmful or injurious outcome |
| Deception | The system is deceptive |
| Distrust | I am suspicious of the system's intent, action, or output |
| Fallacy | The system is deceptive |
| Harm | The system's outcome will have a harmful or injurious outcome |
| Lie | The system is deceptive |
| Misleading | The system is deceptive |
| Mistrust | I am suspicious of the system's intent, action, or output |
| phony | The system is deceptive |
| Sneaky | The system behaves in an underhanded manner |
| Steal | The system behaves in an underhanded manner |
| Suspicion | I am suspicious of the system's intent, action, or output |

We used 80% data for training, 20% data for the test, and 10-fold cross-validation to calculate average accuracy over the folds. We checked for other combinations of train-test data (e.g., 90%-10%), but it did not impact the average accuracy significantly; due to the large size of the dataset, we consider that 80% of training data is good enough to build a model with Machine Learning. We will address the average accuracy as Accuracy from hereafter for the convenience of discussion. We have performed below analysis to check the accuracy:

- Seven classifier algorithms with Machine Learning for each participant for trust and distrust classification and identified the best algorithm for each individual.
- We have also checked this model in the general scenario, which means all the participants' data together. For the convenience of discussion, we will address this all-together dataset as General dataset from hereafter in this discussion.
- We have identified the lesser number (15, 10, and 5) of channels, with the feature selection method using Univariate Statistical Test for each individual as well as for General data and observed model accuracy with those reduced features.
- The best channels identified for General dataset, were applied to the individual datasets to observe the performance.
- We also performed the cleaning of one participant's dataset to observe the effect of the accuracy. The purpose of this study is to develop a model that will provide satisfactory accuracy to sense Trust and Distrust in a real-life scenario with minimum resources, such as computational cost and time. Therefore, we aim to delve deep and explore different instances with raw or uncleaned data, which will use data without any kind of pre-processing directly collected from the user's head and fed to the 'sensing' system. Still, we wanted to examine how much accuracy we are losing by avoiding pre-processing; that is, what is the impact of having all the noises and artifacts in the data. Therefore, we cleaned one participant's data following the steps as shown in Figure 2. We filtered the data with high pass 1Hz and low pass 50 Hz. And while cleaning, as we are already taking the path of pre-processing, we took the chance to observe if averaging over the epochs provides an improvement (or deterioration) of the Accuracy, compared to the Accuracy achieved from individual sample points of every trial.

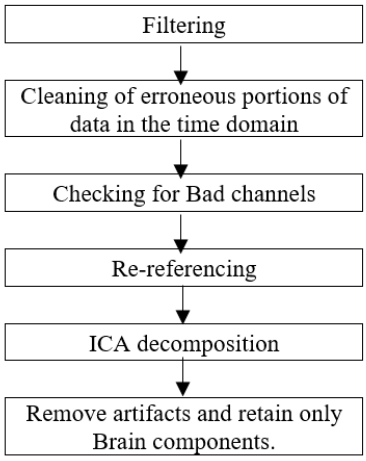



Figure 2. Data cleaning process to retain only the brain component

# 4. Data Collection

## 4.1 Participants

In this study, 5 participants took part with their informed consent as they were explained the process and informed about probable discomfort. IRB approval is acquired for this study, and a consent form is signed by the participants. Among 5 of them, three are male, and two are female by their self-identification. The participants had normal or corrected to normal vision, were in a stable mental condition during the study, were free of past or current neurological or psychological disorders, and were capable of reading the English language properly. The age range of the participants is 18-35 years.

## 4.2 Procedure

The participants were clearly explained with the procedure before starting the experiment. The trust and distrust recording was done separately, each with a preparatory session. The trust session is done at first. The 15 words of trust are introduced to the participants, and they are told to recall events or experiences regarding the words, which must be associated with machines, any computerized system, websites or apps, etc. The cue lines associated with

those words are also shown to them to understand the association with machines. Once they inform that they are ready with their memories related to the words and machines. They are fitted with the EEG head cap and seated in front of the computer screen. Then, the trust words appear in front of them one after another, along with their cue lines, as shown in Figure 3(b). Each word is presented for 30 seconds, and in those 30 seconds, participants are to think of the experience or event associated with that word and machines. Then, a blank screen appears for 5 seconds, followed by the next word. After the trust session, there is a gap of 5 minutes, and then, in a similar pattern, the distrust session is conducted. The reason to conduct the trust session before the distrust session is negative feelings (maybe associated with distrust words) can stay in the human brain more than positive feelings. The gap between the trust session and the distrust session is to reduce any remaining positive feelings evoked in the trust session.

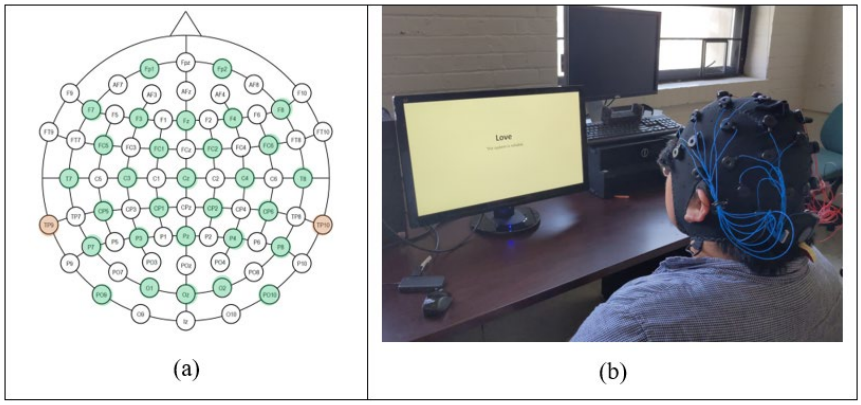



Figure 3. Extended 10-20 electrodes system with highlighted electrodes that are used in this study, (b) A participant is taking part in the study wearing the head-cap

## 5. Results and Discussion

### 5.1 Raw Data

We will discuss the results of the raw data of all instances and cleaned data of one participant in this section. We have calculated the average accuracy of the 10-fold cross-validation training dataset for all of our 5 participants' raw datasets individually as well as for the General dataset, as described in the data analysis section. The results are shown in Table 3. The best Accuracy cells are highlighted in green.

From Table 3, we can notice that for all the scenarios (General data and individual data), the KNN and/or RF algorithm provides the highest Accuracy, which is 100% or close to 100%. With the aim to increase efficiency in terms of data collection, data computation, and user comfort, we explored scenarios with the reduced number of electrodes, that is, 15, 10, and 5 electrodes. We have identified the best 15, 10, and 5 channels for General and individual datasets, as shown in Table 7. In Table 7, the channels are sorted in the order from the highest variance to the lowest variance as obtained from the Univariate analysis. We calculated the Accuracy with the lower number of channels for all instances and presented the result in Tables 4, 5, and 6.

As we identified the best channels of General dataset, we used those channels on the individual dataset to calculate Accuracy. Results of all algorithms are not presented here, as the performance resonates with already-presented results; we found that KNN and/or RF provide the best Accuracy for Genera-best features applied for individual datasets. To show the comparison of performance achieved from General-driven best channels and Individual-driven best channels on individual cases, only the best algorithm (KNN/RF) Accuracies are shown in Figure 4. Therefore, here we can observe, for an individual, let us say for P1, the performance of the model based on his/her own best features as identified and presented in Table 7 column "P1" as well as the performance of the model based on the General best feature as identified and presented in Table 7 column "General."

## 5.2 Clean data

From the overall comparison presented in Figure 4, we can notice that the P3 participant's Accuracy is lower in general than other participants. Therefore, we considered the P3 dataset for pre-processing and cleaning for investigation. As we retain only Brain-majority components, we removed all other components, such as eye, heart, muscle, or other-majority components. By independent component analysis, we obtained 30 components, and labeled them with their maximum variance source. We have removed all other components than those labeled as "Brain" to retain only the data portion having the majority variance coming from Brain. Thus, we have removed a large portion of the original signal. After that, we extracted epochs for Trust and Distrust and performed classification analysis on both Trail and Averaged scenarios. In the Trial scenario, we considered all the individual sample points collected from all the Trials, hence having a total of 115,200 data points. For averaged over epochs, we considered the average of 15 trials for similar time points, hence having a total of 7,580 data points.

Table 3. Classification results of individual participants and General dataset using raw data with all features

| All 30 channels, Average Accuracy in Percentage (%) | | | | | | |
|---|---|---|---|---|---|---|
| | **General** | **P1** | **P2** | **P3** | **P4** | **P5** |
| LR | 59.0 | 59.0 | 69.1 | 77.6 | 69.4 | 64.5 |
| LDA | 58.9 | 58.9 | 68.9 | 76.1 | 69.1 | 63.7 |
| KNN | 99.5 | 99.5 | 99.5 | 99.9 | 100.0 | 98.7 |
| DT | 93.7 | 93.7 | 96.0 | 95.9 | 98.3 | 92.5 |
| NB | 54.5 | 54.5 | 77.4 | 57.0 | 89.8 | 61.3 |
| SVM | 85.9 | 85.9 | 94.5 | 93.4 | 96.2 | 83.5 |
| RF | 98.9 | 99.0 | 99.5 | 99.6 | 97.4 | 98.2 |

Table 4. Classification results from the best 15 channels for all individuals and General dataset

| Best 15 channels, Average Accuracy in Percentage (%) | | | | | | |
|---|---|---|---|---|---|---|
| | **General** | **P1** | **P2** | **P3** | **P4** | **P5** |
| LR | 57.9 | 63.5 | 72.5 | 65.4 | 59.9 | 66.8 |
| LDA | 57.4 | 63.5 | 71.9 | 64.7 | 59.4 | 67.1 |
| KNN | 98.2 | 98.8 | 99.4 | 98.1 | 96.7 | 97.1 |
| DT | 88.5 | 94.1 | 94.8 | 85.0 | 88.2 | 85.9 |
| NB | 54.0 | 74.7 | 56.6 | 59.3 | 58.1 | 69.4 |
| SVM | 76.4 | 91.9 | 87.2 | 78.2 | 75.1 | 86.2 |
| RF | 96.4 | 97.8 | 98.7 | 94.4 | 95.4 | 93.6 |

Table 5. Classification results from the best 10 channels for all individuals and General dataset

| Best 10 channels, Average Accuracy in Percentage (%) | | | | | | |
|---|---|---|---|---|---|---|
| | **General** | **P1** | **P2** | **P3** | **P4** | **P5** |
| LR | 57.9 | 61.9 | 72.3 | 61.4 | 56.8 | 66.6 |
| LDA | 57.7 | 62.1 | 71.6 | 60.5 | 57.0 | 66.5 |
| KNN | 94.6 | 97.0 | 98.1 | 95.0 | 94.9 | 93.0 |
| DT | 83.2 | 89.7 | 93.1 | 81.9 | 86.4 | 82.4 |
| NB | 55.7 | 72.4 | 56.4 | 59.2 | 58.6 | 59.3 |
| SVM | 71.0 | 88.9 | 84.5 | 72.3 | 73.7 | 81.6 |
| RF | 92.3 | 95.8 | 97.3 | 91.6 | 93.2 | 90.1 |

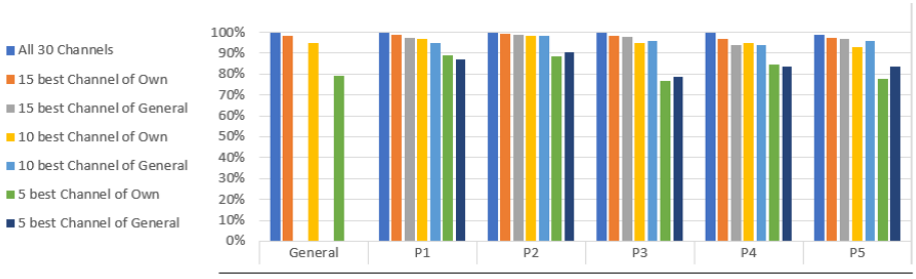
Table 6. Classification results from the best 5 channels for all individuals and General dataset

| Best 5 channels, Average Accuracy in Percentage (%) | | | | | | |
|---|---|---|---|---|---|---|
| | **General** | **P1** | **P2** | **P3** | **P4** | **P5** |
| LR | 56.9 | 61.4 | 69.4 | 60.6 | 54.1 | 65.2 |
| LDA | 56.7 | 61.4 | 69.2 | 59.4 | 54.3 | 65.2 |
| KNN | 79.1 | 89.2 | 88.5 | 76.7 | 84.4 | 76.4 |
| DT | 71.7 | 82.9 | 82.5 | 69.8 | 79.4 | 71.2 |
| NB | 53.9 | 70.2 | 59.3 | 57.0 | 59.4 | 58.0 |
| SVM | 63.9 | 81.5 | 75.0 | 64.6 | 69.3 | 75.9 |
| RF | 79.1 | 88.7 | 88.3 | 80.3 | 84.7 | 77.8 |

Table 7. The best 15 channels for all the 5 participants and general dataset with raw data

| | **General** | **P1** | **P2** | **P3** | **P4** | **P5** |
|---|---|---|---|---|---|---|
| 1 | Fp2 | FC6 | Fp1 | PO9 | Fp2 | CP6 |
| 2 | Fp1 | Fp2 | Fp2 | FC5 | Fp1 | C4 |
| 3 | PO9 | T8 | Cz | CP1 | T8 | Cz |
| 4 | CP1 | Fz | CP1 | T7 | CP5 | Oz |
| 5 | CP5 | F8 | C4 | F7 | F4 | Fz |
| 6 | FC5 | CP6 | Pz | F8 | F8 | F3 |
| 7 | O1 | C4 | FC1 | T8 | F3 | F4 |
| 8 | FC1 | Fp1 | F4 | CP5 | P3 | Fp2 |
| 9 | FC2 | FC5 | Oz | Fp2 | T7 | FC6 |
| 10 | PO10 | F3 | P3 | F3 | PO10 | P4 |
| 11 | P4 | Cz | PO10 | CP2 | CP6 | CP5 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 12 | P8 | P4 | F3 | FC6 | C3 | F8 |
| 13 | F3 | C3 | P4 | Pz | P8 | P3 |
| 14 | Pz | FC2 | O2 | P4 | P7 | C3 |
| 15 | P3 | T7 | FC6 | F4 | C4 | FC2 |




| | General | P1 | P2 | P3 | P4 | P5 |
|---|---|---|---|---|---|---|
| All 30 Channels | 99.5% (KNN) | 99.5% (KNN) | 99.5% (KNN) | 99.9% (KNN) | 100.0% (KNN) | 98.7% (KNN) |
| 15 best Channel of Own | 98.2% (KNN) | 98.8% (KNN) | 99.4% (KNN) | 98.1% (KNN) | 96.7% (KNN) | 97.1% (KNN) |
| 15 best Channel of General | | 97.1% (KNN) | 98.8% (KNN) | 97.6% (KNN) | 94% (KNN) | 96.8% (KNN) |
| 10 best Channel of Own | 94.6% (KNN) | 97.0% (KNN) | 98.1% (KNN) | 95.0% (KNN) | 94.9% (KNN) | 93.0% (KNN) |
| 10 best Channel of General | | 94.6% (KNN) | 98.2% (KNN) | 95.9% (KNN) | 93.6% (KNN) | 95.9% (KNN) |
| 5 best Channel of Own | 79.1% (RF & KNN) | 89.2% (KNN) | 88.5% (KNN) | 76.7% (KNN) | 84.7% (RF) | 77.8% (RF) |
| 5 best Channel of General | | 86.8% (KNN) | 90.6% (RF) | 78.9% (KNN) | 83.5% (KNN) | 83.5% (KNN) |

Figure 4. Comparison of Individual's own best features and General best features

## 5.2 Clean data

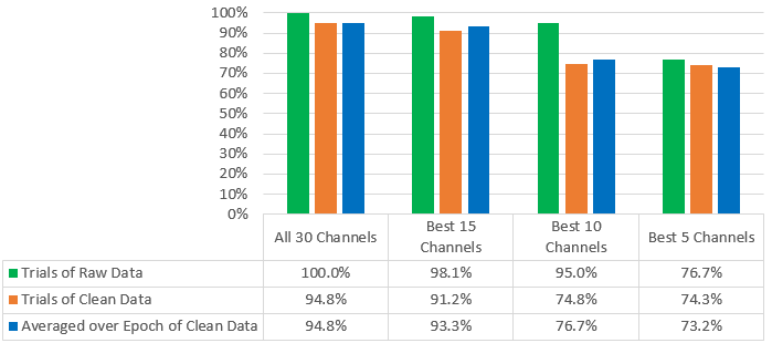
From the overall comparison presented in Figure 4, we can notice that the P3 participant's Accuracy is lower in general than other participants. Therefore, we considered the P3 dataset for pre-processing and cleaning for investigation. As we retain only Brain-majority components, we removed all other components, such as eye, heart, muscle, or other-majority components. By independent component analysis, we obtained 30 components, and labeled them with their maximum variance source. We have removed all other components than those labeled as "Brain" to retain only the data portion having the majority variance coming from Brain. Thus, we have removed a large portion of the original signal. After that, we extracted epochs for Trust and Distrust and performed classification analysis on both Trail and Averaged scenarios. In the Trial scenario, we considered all the individual sample points collected from all the Trials, hence having a total of 115,200 data points. For averaged over epochs, we considered the average of 15 trials for similar time points, hence having a total of 7,580 data points.

Table 8. Cleaned data classification analysis with the P3 dataset

| | **All 30 Channels** | | **Best 15 Channels** | | **Best 10 Channels** | | **Best 5 Channels** | |
|---|---|---|---|---|---|---|---|---|
| | Trial | Avg | Trial | Avg | Trial | Avg | Trial | Avg |
| LR | 50.1% | 46.5% | 50.1% | 46.7% | 50.7% | 45.5% | 50.6% | 46.0% |
| LDA | 50.1% | 46.5% | 50.1% | 46.5% | 50.7% | 45.8% | 50.6% | 46.0% |
| KNN | 92.1% | 93.7% | 89.5% | 92.0% | 68.6% | 71.6% | 71.8% | 73.2% |
| DT | 90.8% | 91.7% | 86.1% | 89.6% | 66.1% | 69.7% | 66.2% | 66.4% |
| NB | 83.9% | 88.9% | 83.2% | 88.4% | 57.2% | 60.4% | 58.7% | 61.4% |
| SVM | 92.2% | 91.7% | 89.0% | 90.9% | 69.7% | 71.1% | 72.4% | 71.6% |
| RF | 94.8% | 94.8% | 91.2% | 93.3% | 74.8% | 76.7% | 74.3% | 73.2% |

We also identified the best channels in a similar way as we did for raw data. We found that the largest variances are coming from the channels F4, Pz, P3, O1, CP2, CP1, FC2, C3, FC1, PO9, C4, Oz, CP6, O2, CP5 as sorted according to the variance. We calculated the Accuracy with the best 15, 10, and 5 channels. While looking at the

averaged dataset, we found that the 15 best channels are the same for averaged and trial datasets, which was expected, though their variance-wise sequence is a little different. We calculated the Accuracy with the best 15, 10, and 5 channels for both the trial and averaged dataset, as shown in Table 8. To visualize the comparison of Trial and Averaged dataset performance, only the best results with different numbers of channels are presented in Figure 5. We can notice that the averaged data result is better than the Trial dataset for this individual with clean data.




| | All 30 Channels | Best 15 Channels | Best 10 Channels | Best 5 Channels |
|---|---|---|---|---|
| Trials of Raw Data | 100.0% | 98.1% | 95.0% | 76.7% |
| Trials of Clean Data | 94.8% | 91.2% | 74.8% | 74.3% |
| Averaged over Epoch of Clean Data | 94.8% | 93.3% | 76.7% | 73.2% |

Figure 5. Classification accuracy best results on Trial and Averaged dataset of P3

### 5.3 Validation

The purpose of developing this model is to train and adjust the machine programs with the individual human operator with the word-elicitation scheme, and then, when in action, the machine program will continuously re-train the model with the practical data. From the results of this preliminary analysis, we can notice that the best classification algorithm is KNN or RF in all the scenarios with raw data. We can also state that KNN is providing unarguably the best result with a larger number of electrodes, whereas RF is providing the best result with a lower number of electrodes. The Decision Tree algorithm is another algorithm which provides quite acceptable accuracy in all the scenarios. The other compatible algorithm is SVM in terms of accuracy, but the training time is multiple folds greater than the other ones, which is not suitable for our live-streaming-based model.

As we observe the comparison of individual and generalistic approaches, we can observe that the individualistic approach provides better accuracy than the generalistic approach. Though we cannot conclude the fact with the small number of participants, we can initiate a hypothesis that an individualistic model provides better accuracy than a 'one-size-fits-all' model. When we examined the scenarios of reduced channels, surprisingly, we noticed that in some cases, the best features obtained from the General dataset provide better accuracy for the classification of the individual dataset, especially in the case of channels; we can see that P2, P3, and P5 has better accuracy with the General best channels than their own best channels. That may be an indication that the Univariate Statistical Test might not be the best method for feature selection; further study can be conducted to address the issue. Even in other cases of reduced scenarios, we can see that the General best channels provide compatible accuracy as an individual's own best channel. From that, we can state that it is possible to develop a general model trained with a large number of participants' data, which will be able to provide satisfactory accuracy for all humans. Here, we need to keep in mind that the satisfactory accuracy might differ from situation to situation. In some cases, more than 80% accuracy might be enough, but in some other cases, we may need accuracy as good as 99%. That is why, here, we are exploring different approaches so that, based on the necessity of the instances, we can decide which way we should take.

In this study, our aim is to calibrate trust with the lowest computational cost while achieving the desired minimum accuracy. That is why we considered raw data, that is, collected brain signals directly feeding for classification analysis without any kind of pre-processing and cleaning. Nevertheless, we wanted to investigate the effect of cleaning on the classification accuracy, so that we can consider that as an option if needed. We can see that the accuracy with clean data is lower than the raw data in all instances for the considered individual. However, if we look at the clean data scenarios, we can see that averaging over epochs can provide better accuracy than individual trial's data points, at least for this specific participant. Also, we can notice that the RF is unarguably the best algorithm for a cleaned dataset. Also, SVM and Decision Tree provide accuracy nearly or even better than KNN. Therefore, we can state that for a situation that uses pre-processed data, and we can include some cleaning, perhaps the RF algorithm will perform better. The improvement of the result with the Naïve Bayes algorithm is also notable.

We can also find that the best channels obtained from the clean data are very different than those of raw data. This can be understood from Figure 2; we can see that by elaborate cleaning and pre-processing, we made changes and even removed selected components, including eye and muscle components, which can have contributed to increasing/decreasing accuracies. We can see that in raw data instances, as mentioned in Table 7, quite a number of best channels are from the Frontal region; probably some of those variances are from the eye, and it is already established by research that eye movements are associated with emotion and cognitive load (Calvo and Nummenmaa 2008; Hyönä and Lorch 2004); which again are associated with trust and distrust. However, in this instance, we only retained Brain-majority signals and removed everything else, as in this study, we just wanted to observe the effect of cleaning. For more specific instances, we can probably re-assess which components we want to remove and how that will affect the accuracy.

From the cleaned data analysis of one participant, we can see that this elaborate process of cleansing is consuming computational resources and not adding any value. However, we cannot conclude that from this preliminary analysis of one participant's data, nor it is our goal of this study. Rather, we can state that this model of word-elicitation study is worth exploring deeply for trust calibration and can lead to establishing a robust model using only brain signals. Using brain signals has its own challenges, such as wearing an EEG headset can cause discomfort and fatigue to the user. Therefore, it can be an issue where the EEG headset can add extra load to the user, but in cases where EEG headsets are already worn (like BCI prosthetics, BCI controlled robotic arm or exoskeletons), the tuning of machine behavior sensing the trust/distrust of the human will be of great advantages. Also, where EEG devices are not originally worn, we can assess the cost-benefit analysis of adding a lighter headset where sensing live trust can provide enough value, as commercial companies are offering low-cost less-discomfortable devices and we can achieve probably-acceptable (acceptable or not, it will depend on situation) accuracy with lower number of electrodes.

### 5.4 Proposed Improvement

From this preliminary analysis of the model, we cannot conclude about the efficacy of the general model, though even with the reduced number of electrodes, with all the participants with their individual datasets, we achieved more than 95% accuracy. Rather, we can consider these findings as a first step to calibrate human-machine trust in a live scenario. Also, we cannot conclude that the individualistic method will work for everyone, as the participant number is not enough for that, but with all the participants' good results, we can be hopeful that the individualistic method will work for most of the individuals if not for all. Therefore, as per the preliminary analysis, we can hypothesize that the general model can help us to identify reduced features, but we will need the individual's own data to train the model for better accuracy. Also, we cannot fix whether KNN or RF would be better for raw data classification. In the future, we plan to extend this study with more participants to make the generalistic model robust as well as to confirm the reduced channels which will be applicable to everyone. We also plan to delve more into the data and explore other processes to increase the efficiency and accuracy of the model.

We also aim to conduct a more elaborate study with the different data-cleaning methods because, as discussed before, some situations might need some data cleaning, which was not the scope of this presented study; here, we aimed to develop a basic or general approach, more to check the viability of the method through preliminary analysis. To develop the model for a practical scenario, we will need to narrow down our scope to specify our target situation. Only the situation can decide whether we should go for a generalistic approach or individualistic approach with raw data, or whether we need some data cleaning while accepting some computational cost and accuracy loss, or how many channels we want to use.

## 6. Conclusion

In brief, we can state that this article introduces an approach that has a high potential to be used in a real-life scenario where live sensing of trust is of high advantage. With the results of preliminary data analysis, we can conclude that there is a promising possibility to use this method for calibration of trust in different situations.

## References

Akash, K., Hu, W., Jain, N. and Reid, T., A classification model for sensing human trust in machines using EEG and GSR, *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 8, no. 4, pp. 1-20, 2018.

Barber, B. *The logic and limits of trust*, 1st Edition, Rutgers University Press, 1983.

Boudreau, C., McCubbins, M. D. and Coulson, S., Knowing when to trust others: An ERP study of decision making after receiving information from unknown people, *Social Cognitive and Affective Neuroscience*, vol. 4, no. 1, pp. 23-34, 2009.

Calvo, M. G., and Nummenmaa, L., Detection of emotional faces: Salient physical features guide effective visual search, *Journal of Experimental Psychology: General*, vol. 137, no. 3, 471–494.

Dong, S., Kim, B., Lee, K. and Lee, S., A preliminary study on human trust measurements by eeg for human-machine interactions, *Proceedings of the 3rd International Conference on Human-Agent Interaction*, pp. 265-268, Daegu Kyungpook, Republic of Korea, October 21-24, 2015.

Firoz, K. F. and Seong, Y., A neural study of intuitive mode of cognition while decision-making using artificial grammar learning paradigm, *IISE Annual Conference and Expo*, New Orleans, USA, May 20-23, 2023.

Firoz, K. F. and Seong, Y., A Preliminary Study of Neural Correspondence to Intuitive Binary Decision through Electroencephalography. 2023. Available at SSRN 4579115.

Firoz, K. F., Seong, Y. and Oh, S., A neurological approach to classify trust through EEG signals using machine learning techniques, *2022 IEEE 3rd International Conference on Human-Machine Systems (ICHMS)*, pp. 1-6, Orlando, FL, USA, November 17-19, 2022.

Hoff, K. A. and Bashir, M., Trust in automation: Integrating empirical evidence on factors that influence trust, *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 57, no. 3, pp. 407-434, 2015.

Hyönä, J., and Lorch, R. F., Effects of topic headings on text processing: Evidence from adult readers' eye fixation patterns, *Learning and Instruction*, vol. 14, no. 2, pp. 131-152, 2004.

Jian, J., Bisantz, A. M., and Drury, C. G., Foundations for an empirically determined scale of trust in automated systems, International Journal of Cognitive Ergonomics, vol. 4, no. 1, pp. 53-71, 2000, doi:10.1207/S15327566IJCE0401_04.

Merritt, S. M., and Ilgen, D. R., Not all trust is created equal: Dispositional and history-based trust in human-automation interactions, *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 50, no. 2, pp. 194-210, 2008.

Muir, B. M., Trust between humans and machines, and the design of decision aids, *International Journal of Man-Machine Studies*, vol. 27, no. 5-6, pp. 527-539, 1987.

Oh, S., Seong, Y., Yi, S., and Park, S., Neurological measurement of human trust in automation using electroencephalogram, *International Journal of Fuzzy Logic and Intelligent Systems*, vol. 20, no. 4, pp. 261-271, 2020.

Parasuraman, R., and Riley, V., Humans and automation: Use, misuse, disuse, abuse, *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 39, no. 2, pp. 230-253, 1997, doi:10.1518/001872097778543886.

Sheridan, T. B., Trustworthiness of command and control systems. *Third IFAC/IFIP/IEA/IFORS Conference*, Oulu, Finland, June 14-16, 1989, pp. 427-431.

## Acknowledgments

The authors would like to acknowledge support from the NSF Engineering Research Center for Hybrid Autonomous Manufacturing Moving from Evolution to Revolution (ERC-HAMMER) under Award Number EEC-2133630.

## Biographies

**Kazi Farzana Firoz** is a Ph.D student in the Industrial and Systems Engineering department at North Carolina A&T State University. Her research area is human-machine systems with a focus on neuro-ergonomics, where she works on exploring human brain signals for the advancement of human-machine interaction. For her research, she uses Digital Signal Processing, Machine Learning, and other advanced techniques. She completed her B.Sc. in Industrial and Production Engineering from Bangladesh University of Engineering and Technology (BUET).

**Younho Seong** is a professor in the Industrial and Systems Engineering department at North Carolina A&T State University. His research interests include judgment and decision-making, human trust in automation, human-computer interface, data visualization, neuroergonomics, and brain-computer interface. He completed his Ph.D in Industrial Engineering from The State University of New York at Buffalo. He completed his MSE in Industrial Engineering and B.S. in Industrial and Systems Engineering from Inha University.

**Sun Yi** is a professor in the Mechanical Engineering department at North Carolina A&T State University. His research interests include analysis and control of dynamic systems with application to robots, vehicles, and aircraft.

He completed his Ph.D and MS. in Mechanical Engineering from The University of Michigan. He completed his B.S. in Mechanical and Aerospace Engineering from Seoul National University.

**Yoo-Sang Chang** is a Ph.D student in the Industrial and Systems Engineering department at North Carolina A&T State University. His research area is human-machine system, human's decision-making, and neuroergonomics. He completed M.E. in Industrial & Management Engineering and B.S.E. in Industrial Management Engineering from Hanbat National University.